Asymptotic Mean Stationarity of Sources With Finite Evolution Dimension

Ulrich Faigle and Alexander Schönhuth

Mathematisches Institut
Zentrum für Angewandte Informatik
Universität zu Köln
Weyertal 80, 50931 Köln, Germany
faigle@zpr.uni-koeln.de
aschoen@zpr.uni-koeln.de

Abstract. The notion of the *evolution* of a discrete random source with finite alphabet is introduced and its behavior under the action of an associated linear *evolution operator* is studied. Viewing these sources as possibly stable dynamical systems it is proved that all random sources with finite evolution dimension are asymptotically mean stationary, which implies that such random sources have ergodic properties and a well-defined entropy rate. It is shown that the class of random sources with finite evolution dimension properly generalizes the well-studied class of finitary stochastic processes, which includes (hidden) Markov sources as special cases.

Keywords. Asymptotic mean, dimension, entropy, ergodic, evolution operator, hidden Markov model, linearly dependent process, Markov chain, observable operator model, random source, stable, state generating function, stationary

1 Introduction

A central problem of data analysis is learning from sequences that appear to be issued by a random source. In order to admit appropriate learning models, however, the random source should be such that sampling yields reliable information. As pointed out in Choi *et al.* [3], for example, most models simply go on the assumption that the random source in question is stationary, which typically is *not* the case–even when the source is Markov. Moreover, also the theoretical literature usually restricts the study of ergodic and entropic properties to stationary random sources (see, *e.g.*, Han and Kobayashi [7]).

Birkhoff's *ergodic theorem*, on the other hand, provides a key to a certain converse. One can show that the presence of ergodic properties with respect to bounded measurements is equivalent to the seemingly weaker property of asymptotic mean stationarity. Moreover, asymptotically mean stationary sources guarantee the *entropy ergodic theorem* of Shannon-McMillan-Breiman to hold, see [13].

Therefore, it is of both theoretical and practical interest to know which random sources are asymptotically mean stationary (AMS). It is the purpose of the present note to exhibit a large class of random sources to be AMS. We show that this class

properly contains the class of so-called *finitary* (a term coined by Heller, see [8]) or *linearly dependent processes* (LDPs) (see Ito *et al.* [10]), which arose from the attempt to understand hidden Markov models (HMMs) within the framework of linear algebra (see Gilbert [6] and Heller [8],[9]). Moreover, as LDPs allow finite parametrizations, they offer a promising model for the construction of general learning algorithms (see Jaeger [12], who studies them as *concrete observable operator models*). It is known that LDPs properly generalize HMMs (Heller [8], Jaeger [12]).

We define and study our class of (one-sided) random sources by identifying *ground states* of arbitrary discrete random sources and analyzing their behavior under the action of a (linear) *evolution operator*. HMMs have been shown to be AMS by Kieffer and Rahe [14]. We prove more generally that all finite alphabet discrete random sources with finite *evolution dimension* are necessarily AMS (Corollary 3). In our model, a source is stationary exactly when its evolution dimension equals 1.

Our approach views discrete random sources as dynamical systems that evolve under the action of linear operators. The asymptotic mean stationarity then translates into the existence of the Cesàro average for the evolution operator. In this context, our main result is Theorem 2, which characterizes the finite-dimensional linear operators with Cesàro property as the *stable* operators in the sense of Brayton and Tong [2].

2 States and Evolution

As usual, Σ^* denotes the set of all strings of finite length over the finite alphabet Σ together with the concatenation operation:

$$w \in \Sigma^t, v \in \Sigma^k \quad \Longrightarrow \quad wv \in \Sigma^{t+k} \ .$$

We write $|w| = t$ for the *length* of $w \in \Sigma^t$. We think of the random source (X_t) as being specified by a function

$$p : \Sigma^* \to [0,1] \subseteq \mathbb{R} \quad \text{such that} \quad \sum_{a \in \Sigma} p(wa) = p(w) \quad \text{for all } w \in \Sigma^*, \tag{1}$$

assuming $p(\square) = 1$, where the word $\square \in \Sigma^0$ of length $|\square| = 0$ is the *empty string*, which implies

$$\sum_{w \in \Sigma^t} p(w) = 1 \quad \text{for all } t = 0, 1, \ldots. \tag{2}$$

Note that these functions describe the class of one-sided random processes with values in Σ. For sake of technical simplicity, we imagine that the random source (X_t) under consideration emits the empty string $X_0 = \square$ at time $t = 0$ and, at time $t \geq 1$, will have produced a string $w = w_1 w_2 \ldots w_t \in \Sigma^t$ with the probability

$$p(w) = \Pr\{X_1 = w_1, X_2 = w_2, \ldots, X_t = w_t\} \ .$$

The condition $p(\square) = 1$ can then be translated to that the probability of emitting an arbitrary sequence is one. Given $w \in \Sigma^t$, the probability to obtain the string $v =$

$v_1 \dots v_k \in \Sigma^k$ in the next k time periods is

$$p(v|w) = \Pr\{X_{t+1} = v_1, \dots, X_{t+k} = v_k | w\} = \begin{cases} 0 & \text{if } p(w) = 0 \\ p(wv)/p(w) & \text{if } p(w) \neq 0, \end{cases}$$

namely the corresponding (conditional) *prediction probability*, and we have

$$\sum_{v \in \Sigma^k} p(v|w) = 1 \quad \text{whenever } p(w) \neq 0.$$

Upon having seen the string w at time t, we think of the random source as being in a *state* that depends only on w and completely describes the probabilities for the symbols to be produced at time $t+1$. All this information is contained in the (infinite) vector of prediction probabilities

$$\mathbf{g}_w = [p(v|w)_{v \in \Sigma^*}] \in \mathbb{R}^{\mathbb{N}} ,$$

which suggests to identify the possible states with these vectors $\mathbf{g}_w$. We collect these vectors as columns into the (non-negative) *prediction matrix*

$$\mathcal{P} = [p(v|w)_{v,w \in \Sigma^*}] .$$

2.1 The Prediction Space and the State Space

The *prediction space* $\mathcal{V}$ of (X_t) is defined as the column space of $\mathcal{P}$, *i.e.* the set of all (finite) linear combinations of states $\mathbf{g}_w$. The *state space* $\mathcal{S}$ is the affine subspace of those vectors $\mathbf{v} \in \mathcal{V}$ whose components $\mathbf{v}(a), a \in \Sigma$, add up to 1:

$$\mathcal{S} = \{\mathbf{v} \in \mathcal{V} | \sum_{a \in \Sigma} \mathbf{v}(a) = 1\} .$$

Since $\mathcal{S}$ contains all state vectors $\mathbf{g}_w$, the next observation is obvious.

Lemma 1. *Let $\mathbf{g}_{w_1}, \dots, \mathbf{g}_{w_k}$ be arbitrary state vectors and $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ scalars. Then*

$$\mathbf{v} = \sum_{i=1}^{k} \alpha_i \mathbf{g}_{w_i} \in \mathcal{S} \quad \Longleftrightarrow \quad \sum_{i=1}^{k} \alpha_i = 1 .$$

◇

The random source (X_t) is in the state $\mathbf{g}^0 = \mathbf{g}_\square$ at time $t = 0$. The expected state at time $t \geq 1$ is the so-called tth *ground state*

$$\mathbf{g}^t := \sum_{w \in \Sigma^t} p(w) \mathbf{g}_w \ \in \ \mathcal{S} .$$

Note that the component of $\mathbf{g}^t$ corresponding to $v = v_1 \dots v_k$ is

$$\mathbf{g}^t(v) = \sum_{w \in \Sigma^t} p(w) p(v|w) = \Pr\{X_{t+1} = v_1, \dots, X_{t+k} = v_k\} .$$

Hence the discrete random source (X_t) is stationary if and only if there exists only one ground state:

$$\mathbf{g}^0 = \mathbf{g}^1 = \ldots = \mathbf{g}^t = \ldots .$$

2.2 The Evolution Operator

Given (X_t) is in the state $\mathbf{g}_w \neq 0$, the expected next state is

$$\psi(\mathbf{g}_w) = \sum_{a\in\Sigma} p(a|w)\mathbf{g}_{wa} \in \mathcal{S} \ .$$

We extend ψ to a linear operator on $\mathcal{V}$ *via*

$$\psi(\sum_{i=1}^{k} \alpha_i \mathbf{g}_{w_i}) := \sum_{i=1}^{k} \alpha_i \psi(\mathbf{g}_{w_i}) \ .$$

Lemma 2. *$\psi : \mathcal{V} \to \mathcal{V}$ is a well-defined linear operator.*

Proof. Assume $\sum_{i=1}^{k} \alpha_i \mathbf{g}_{w_i} = \mathbf{g}_w$ for some $w \in \Sigma^*$. Then we find for all $v \in \Sigma^*$,

$$\begin{aligned}\psi(\mathbf{g}_w)(v) = \sum_{a\in\Sigma} p(a|w)p(v|wa) \ &= \ \sum_{a\in\Sigma}\sum_{i=1}^{k} \alpha_i p(av|w_i) \\ = \sum_{i=1}^{k} \alpha_i \sum_{a\in\Sigma} p(av|w_i) \ &= \ \sum_{i=1}^{k} \alpha_i \psi(\mathbf{g}_{w_i})(v) \ .\end{aligned}$$

◇

We call ψ the *evolution operator* of (X_t). In particular, we note that the ground states evolve from the initial state $\mathbf{g}^0$ by successive applications of ψ:

$$\begin{aligned}\psi(\mathbf{g}^t) = \sum_{w\in\Sigma^t} p(w)\psi(\mathbf{g}_w) \ &= \ \sum_{w\in\Sigma^t}\sum_{a\in\Sigma} p(a|w)p(w)\mathbf{g}_{wa} \\ = \sum_{w\in\Sigma^t}\sum_{a\in\Sigma} p(wa)\mathbf{g}_{wa} \ &= \ \sum_{v\in\Sigma^{t+1}} p(v)\pi_v \ = \ \mathbf{g}^{t+1} \ ,\end{aligned}$$

which implies

$$\mathbf{g}^{t+1} = \psi(\mathbf{g}^t) = \psi^2(\mathbf{g}^{t-1}) = \ldots = \psi^{t+1}(\mathbf{g}^0) \ .$$

2.3 The State Generating Function

The evolution operator $\psi : \mathcal{V} \to \mathcal{V}$ naturally decomposes into a sum of operators

$$\sigma^a : \mathcal{V} \to \mathcal{V} \quad via \quad \sigma^a(\mathbf{g}_w) := p(a|w)\mathbf{g}_{wa}.$$

In the same way as with ψ, it is straightforward to check that σ^a is indeed a well-defined linear operator. Multinomial expansion of ψ results in the representation

$$\psi^t(\mathbf{g}^0) = (\sum_{a\in\Sigma} \tau^a)^t \mathbf{g}^0 = \sum_{w\in\Sigma^t} \sigma^w(\mathbf{g}^0) \ , \tag{3}$$

where we set $\sigma^{a_1 \dots a_t} := \tau^{a_t} \circ \ldots \circ \sigma^{a_1}$. Note that we have, by definition,

$$\sigma^a(\mathbf{g}^0) = p(a|\Box)\mathbf{g}_a = p(a)\mathbf{g}_a \quad \text{for all } a \in \Sigma$$

and hence

$$\sigma^{a_1 \dots a_t}\mathbf{g}^0 = \sigma^{a_t} p(a_1 \dots a_{t-1})\mathbf{g}_{a_1 \dots a_{t-1}} = p(a_1 \dots a_{t-1}a_t)\mathbf{g}_{a_1 \dots a_{t-1}a_t}.$$

Consider now the formal power series

$$\Psi = \frac{1}{1-\psi} = \sum_{t=0}^{\infty} \psi^t = id + \psi + \ldots + \psi^t + \ldots$$

The multinomial expansion (3) of ψ^t now shows that Ψ may be viewed as the *(probability) state generating function* of the discrete random source (X_t).

REMARK. Note that the presented linear operators do actually not transform the original process, but rather generate conditional probabilities of the process distribution. However, the conditional probabilities $\mathbf{g}_w$ establish process distributions themselves, as also for them the relationships from (1) and (2) apply.

2.4 Evolution Space and Dimension

Consider the set of ground states

$$\mathcal{G} = \{\mathbf{g}^0, \dots, \mathbf{g}^t, \dots\} = \{\psi^0\mathbf{g}^0, \dots, \psi^t\mathbf{g}^0, \dots\}$$

and define the *evolution space* $\mathcal{E} \subseteq \mathcal{V}$ as the collection of all linear combinations of the ground states in $\mathcal{G}$. The *evolution dimension* of (X_t) is the linear dimension of $\mathcal{E}$, *i.e.*,

$$e\dim(X_t) := \dim \mathcal{E} \ .$$

If $e\dim(X_t)$ is finite, there is a minimal d such that scalars $c_0, \dots, c_{d-1} \in \mathbb{R}$ exists with the property

$$\mathbf{g}^d = c_0\mathbf{g}^0 + \ldots + c_{d-1}\mathbf{g}^{d-1} \ .$$

By the minimality of d, the set $\mathcal{G}_d = \{\mathbf{g}^0, \dots, \mathbf{g}^{d-1}\}$ is linearly independent. Moreover, the relation

$$\mathbf{g}^{d+m} = \psi^m\mathbf{g}^d = \sum_{i=0}^{d-1} c_i\psi^m\mathbf{g}^i = \sum_{i=0}^{d-1} c_i\mathbf{g}^{i+m} \quad \text{for all } \ m = 1, 2, \dots$$

successively shows that $\mathcal{G}_t$ is a basis for $\mathcal{E}$ and $d = e\dim(X_t)$. We refer to $\mathcal{G}_d$ as the *evolution basis* of (X_t). ψ acts a a linear operator on $\mathcal{E}$. Relative to the basis $\mathcal{G}_d$, the operator $\psi : \mathcal{E} \to \mathcal{E}$ is described by the *evolution matrix*

$$E = \begin{bmatrix} 0 & 0 & \dots & 0 & c_0 \\ 1 & 0 & \dots & 0 & c_1 \\ 0 & 1 & \dots & 0 & c_2 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & c_{d-1} \end{bmatrix} \ \in \ \mathbb{R}^{d\times d} \ .$$

with *characteristic polynomial*

$$\lambda^d = c_0 + c_1\lambda + \ldots + c_{d-1}\lambda^{d-1} \ .$$

Because of $c_0 + \ldots + c_{d-1} = 1$ (*cf.* Lemma 1), we immediately see that E has the eigenvalue $\lambda = 1$.

REMARK. Note that we do not claim $\mathcal{G}_d$ to be a basis for the whole prediction space $\mathcal{V}$. The evolution matrix E has column sums 1, but contains possibly negative coefficients c_i. In this sense, E could be viewed as a Markov transition matrix with "negative transition probabilities" (*cf. Section 4).*

Our main result is that random sources with finite evolution dimension are *asymptotically stable* in the following sense:

Theorem 1. *If the random source (X_t) has finite evolution dimension $d = e\dim(X_t)$, then the associated evolution matrix E possesses the Cesàro average*

$$\overline{E} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} E^k \ \in \ \mathbb{R}^{d\times d}.$$

REMARK. If formulated with the evolution operator on the possibly infinite-dimensional evolution space, the converse of this theorem does not hold. In fact, it is true that the Cesàro average property of the evolution operator is equivalent to that the process is AMS, see [15] for details. See also the discussion subsequent to proposition 1 which gives an example of an AMS process with infinite evolution dimension.

We defer the proof of Theorem 1 to the next section (*cf.* the proof of Theorem 3) and end this section by mentioning some special cases of particular interest.

If $\psi\mathbf{g}^0 = \mathbf{g}^0$ holds, the random source (X_t) is *stationary*. In other words, we have

$$(X_t) \quad \text{stationary} \quad \Longleftrightarrow \quad e\dim(X_t) = 1 \ .$$

Slightly more generally, a random source (X_t) is said to be N*-stationary* if the process $(Y_t) = (X_{tN}, ..., X_{t+2N-1})$ with alphabet Σ^N is stationary. In our language N-stationarity is equivalent to $\psi^N\mathbf{g}^0 = \mathbf{g}^0$. Hence we find:

$$(X_t) \quad N\text{-stationary} \quad \Longrightarrow \quad e\dim(X_t) \leq N \ .$$

Further examples of processes with finite evolution dimension arise from the *quantum random walks* in the sense of Aharonov *et al.* [1]), which may be generalized to *quantum Markov chains* (*cf.* [5] for a preliminary report on the latter. See also Section 4 for an alternative modeling framework for classical finitary processes and a discussion of its relationship with processes of finite evolution dimension).

3 Stability

It is convenient to discuss stability of linear operators in the context of complex vector spaces. Let thus V be a finite-dimensional vector space over the field $\mathbb{C}$ of complex numbers with a norm

$$\mathbf{v} \;\rightarrow\; \|\mathbf{v}\| \;.$$

Let furthermore $F : V \rightarrow V$ be an arbitrary linear operator. We call F *stable* if for all $\mathbf{v} \in V$ there exists some $c = c(\mathbf{v}) \in \mathbb{R}$ with the property

$$\|F^k\mathbf{v}\| \leq c \quad \text{for all} \;\; k = 1, 2, \ldots.$$

REMARK. A stable linear operator F is also "stable" in the sense of Brayton and Tong [2].

We call the linear operator $F : V \rightarrow V$ *asymptotically stable* if, for every $\mathbf{v} \in V$, the following Cesàro average exists:

$$\overline{\mathbf{v}} = \lim_{t\rightarrow\infty} \frac{1}{t} \sum_{k=0}^{t-1} F^k\mathbf{v} \;.$$

We now show that the two stability concepts coincide.

Theorem 2. *Let $F : V \rightarrow V$ be a linear operator. Then F is stable if and only if F is asymptotically stable.*

Proof. We argue by induction on the dimension $n = \dim V$ and assume the Theorem to be true for all vector spaces of dimension $n' < n$. We first note that we may assume without loss of generality that F has exactly one eigenvalue λ.

Indeed, if there are eigenvalues $\lambda_1 \neq \lambda_2$, V admits a direct sum decomposition into two non-trivial F-invariant subspaces V_1 and V_2:

$$V = V_1 \oplus V_2 \quad \text{and} \quad F : V_i \rightarrow V_i \quad (i = 1, 2).$$

Writing $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$ with $\mathbf{v}_i \in V_i$, the linearity of F makes it clear that F is (asymptotically) stable with respect to V if and only if F is (asymptotically) stable with respect to both V_1 and V_2. So the Theorem follows from the induction hypothesis.

Let now λ be the unique eigenvalue of F. For a corresponding eigenvector $\mathbf{v}$ we then find

$$\frac{1}{t}\sum_{k=0}^{t-1} F^k\mathbf{v} = \frac{1}{t}\sum_{k=0}^{t-1} \lambda^k\mathbf{v} = \begin{cases} \mathbf{v} & \text{if } \lambda = 1, \\ \dfrac{\lambda^t - 1}{t(\lambda - 1)}\mathbf{v} & \text{if } \lambda \neq 1, \end{cases} \tag{4}$$

which makes it clear in the case $|\lambda| > 1$ that F is neither stable nor asymptotically stable. So we can assume $|\lambda| \leq 1$ without loss of generality. The Cayley-Hamilton Theorem guarantees the existence of a minimal $m \in \mathbb{N}$ such that

$$(F - \lambda I)^m\mathbf{v} = \mathbf{0} \quad \text{for all } \mathbf{v} \in V,$$

which yields the binomial expansion

$$F^k\mathbf{v} = [(F-\lambda I)+\lambda I]^k\mathbf{v} = \sum_{j=0}^{k}\binom{k}{j}(F-\lambda I)^j\lambda^{k-j}\mathbf{v} = \sum_{j=0}^{m-1}\binom{k}{j}\lambda^{k-j}(F-\lambda I)^j\mathbf{v}\ .$$

In the case $|\lambda|<1$ we have $\binom{k}{j}\lambda^{k-j} \leq k^j\lambda^{k-j} \to 0$ for all k, and hence observe (asymptotic) stability:

$$\lim_{t\to\infty} F^t\mathbf{v} = \mathbf{0} = \lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}F^k\mathbf{v}\ .$$

It remains to analyze the case $|\lambda|=1$. If $m=1$, we have $F=\lambda I$. Because $|\lambda|=1$, F is trivially stable. Moreover, the geometric sum (4) shows that F is also asymptotically stable. For example, if $\lambda\neq 1$, we find

$$\|\frac{1}{t}\sum_{k=0}^{t-1}F^k\mathbf{v}\| \ = \ \|\frac{1}{t}\sum_{k=0}^{t-1}\lambda^k\mathbf{v}\| \ \leq \ \frac{2}{t|1-\lambda|} \ \to \ 0\ .$$

To complete the proof, we show that the operator F is neither stable nor asymptotically stable if $m\geq 2$. To this end, we select some $\mathbf{v}\in V$ such that

$$\mathbf{w} = (F-\lambda I)\mathbf{v} \neq \mathbf{0} \quad \text{and} \quad (F-\lambda I)^2\mathbf{v} = \mathbf{0}\ .$$

The binomial expansion now takes the form

$$F^k\mathbf{v} = \lambda^k\mathbf{v} + k\lambda^{k-1}(F-\lambda I)\mathbf{v} = \lambda^k\mathbf{v} + k\lambda^{k-1}\mathbf{w}\ .$$

So the triangle inequality $\|F^k\mathbf{v}\| \geq k\|\mathbf{w}\| - \|\mathbf{v}\|$ exhibits F as not stable. F cannot be asymptotically stable either. In the case $\lambda=1$, we namely find

$$\frac{1}{t}\sum_{k=0}^{t-1}F^k\mathbf{v} = \mathbf{v} + \frac{1+\ldots+t-1}{t}\mathbf{w} = \mathbf{v}+\frac{t-1}{2}\mathbf{w}\ ,$$

which does not converge. In the case $\lambda\neq 1$, we consider the function

$$f_t(x) = \frac{1-x^t}{t(1-x)} = \frac{1}{t}\sum_{k=0}^{t-1}x^k \quad (x\neq 1)$$

and observe from the binomial expansion above:

$$\frac{1}{t}\sum_{k=0}^{t-1}F^k\mathbf{v} = f_t(\lambda)\mathbf{v} + f_t'(\lambda)\mathbf{w} \quad (\text{if } \lambda\neq 1).$$

It is straightforward to check that $\lim_{t\to\infty} f_t(\lambda)$ exists while $\lim_{t\to\infty} f_t'(\lambda)$ does not exist if $|\lambda|=1$. So F is not asymptotically stable. ◇

Corollary 1. *The linear operator $F: V\to V$ is stable if and only if its Cesàro average*

$$\overline{F} = \lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}F^k$$

exists. ◇

Let us call the linear operator $F : V \to V$ *regular* if

$$F^\infty = \lim_{k\to\infty} F^k \quad \text{exists.}$$

Clearly, regularity of F implies that F is (asymptotically) stable. Regular linear operators admit a characterization in terms of eigenvalues.

Corollary 2. *The linear operator $F : V \to V$ is regular if and only if every eigenvalue λ of F satisfies*

$$\textit{either} \quad \lambda = 1 \quad \textit{or} \quad |\lambda| < 1 \; .$$

Proof. Straightforward along the lines of the proof of Theorem 2. ◇

3.1 Asymptotic Mean Stationarity of Random Sources

The (not necessarily finite-dimensional) random source (X_t) with finite alphabet Σ is said to be *pointwise asymptotically mean stationary* (or to be PAMS, for short) if for all $v = v_1 \dots v_k \in \Sigma^*$ the following averages converge:

$$\frac{1}{t}\sum_{m=0}^{t-1} \Pr\{X_{m+1} = v_1, \dots, X_{m+k} = v_k\} = \frac{1}{t}\sum_{m=0}^{t-1} \mathbf{g}^m(v) \quad \to \quad \overline{\mathbf{g}}(v) \; .$$

We call the corresponding limit function $\overline{\mathbf{g}} : \Sigma^* \to \mathbb{R}$ the *Cesaro average* of the $\mathbf{g}^t$.

Lemma 3. *Assume that the Cesàro average $\overline{\mathbf{g}}$ of the ground states $\mathbf{g}^t$ of (X_t) exists. Then $\overline{\mathbf{g}}$ is the ground state of a stationary discrete random source $(\overline{X}_t)$ with alphabet Σ, called the* sampling limit *of (X_t).*

Proof. Clearly $\overline{\mathbf{g}}(\square) = 1$ and $\overline{\mathbf{g}}(w) \geq 0$ for all $w \in \Sigma^*$. We next observe

$$\begin{aligned}\sum_{a\in\Sigma} \overline{\mathbf{g}}(wa) &= \sum_{a\in\Sigma} \lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} \mathbf{g}^t(wa) \;=\; \lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1}\sum_{a\in\Sigma} \mathbf{g}^t(wa) \\ &= \lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} \mathbf{g}^t(w) \;=\; \overline{\mathbf{g}}(w) \; .\end{aligned}$$

So $\overline{\mathbf{g}}$ describes a discrete random source $(\overline{X}_t)$ with alphabet Σ. To check that $(\overline{X})_t$ is stationary, *i.e.*, that $\sum_{a\in\Sigma} \overline{p}(a)\overline{\mathbf{g}}_a = \overline{\mathbf{g}}$ holds, we consider an arbitrary component $\overline{\mathbf{g}}(v)$ and find

$$\begin{aligned}\sum_{a\in\Sigma} \overline{p}(a)\overline{\mathbf{g}}_a(v) &= \sum_{a\in\Sigma} \overline{p}(a)\overline{p}(v|a) \;=\; \sum_{a\in\Sigma} \overline{\mathbf{g}}(av) \\ &= \lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1}\sum_{a\in\Sigma} \mathbf{g}^k(av) \;=\; \lim_{t\to\infty} \frac{1}{t}\sum_{k=0}^{t-1} \mathbf{g}^{k+1}(v) \;=\; \overline{\mathbf{g}}(v) \; .\end{aligned}$$

◇

Theorem 3. *Every discrete random source (X_t) with finite alphabet Σ and finite evolution dimension $d = e\dim(X_t)$ is pointwise asymptotically mean stationary.*

Proof. By our hypothesis, the matrix $\mathbf{G} = [\mathbf{g}^0, \ldots, \mathbf{g}^{d-1}]$ contains an $(d \times d)$-submatrix $\mathbf{B} = [\mathbf{b}^0, \ldots, \mathbf{b}^{d-1}]$ of full rank d and the evolution space $\mathcal{E}$ is naturally isomorphic to the column space V of $\mathbf{B}$. Moreover, the evolution operator acts as a linear operator $\psi : V \to V$. Since all the components of ground states lie in the interval $[0, 1] \subseteq \mathbb{R}$, we have

$$\|\psi^t \mathbf{b}^i\| \leq \sqrt{d} \quad \text{for all } t \text{ and all } i.$$

Consequently, we find for every linear combination $\mathbf{v} = \sum_i \alpha_i \mathbf{b}^i$,

$$\|\psi^t \mathbf{v}\| \;\leq\; \sum_{i=0}^{d-1} |\alpha_i| \, \|\psi^t \mathbf{b}^i\| \;\leq\; \sqrt{d} \sum_{i=0}^{d-1} |\alpha_i| \quad \text{for all } t \;,$$

which means that ψ is stable on V (Theorem 2). If we represent the elements of V in coordinates relative to $\mathbf{B}$, ψ is described by the evolution matrix E of (X_t), which thus is recognized to be stable as well with Cesàro average

$$\overline{E} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} E^k \;.$$

Each ground state $\mathbf{g}^t$ has unique coordinates $\mathbf{w}^t \in \mathbb{R}^d$ (*i.e.*, $\mathbf{g}^t = \mathbf{G}\mathbf{w}^t$) relative to the evolution basis $\mathcal{G}_d$ of (X_t) with the Cesàro average

$$\overline{\mathbf{w}} = \overline{E}\mathbf{w}^0 = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} E^k \mathbf{w}^0 = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \mathbf{w}^k \;.$$

Consider now any $v \in \Sigma^*$ and denote by Π_v the projection operator $\mathbf{g}^t \to \Pi_v \mathbf{g}^t = \mathbf{g}^t(v)$. Because $\Pi_v \mathbf{G}$ is a linear functional on the coordinate space $\mathbb{R}^d$, we conclude that

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \mathbf{g}^k(v) = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \Pi_v \mathbf{G}\mathbf{w}^k = \lim_{t\to\infty} \Pi_v \mathbf{G} \sum_{k=0}^{t-1} \frac{1}{t} \mathbf{w}^k = \Pi_v \mathbf{G}\overline{\mathbf{w}} = \overline{\mathbf{g}}(v)$$

exists, which was to be shown.

◇

We point to the difference between the PAMS property and the usual property of asymptotically mean stationarity (AMS), which refers to the convergence of the measures associated with the averages $t^{-1} \sum_{k=0}^{t-1} \mathbf{g}^k$ to the measure associated with $\overline{\mathbf{g}}$ on all elements of the σ-algebra generated by the cylinder sets (which are in 1-1-correspondence with subsets of words). In general, one cannot infer the AMS from PAMS. For processes of finite evolution dimension, however, those two notions coincide.

Corollary 3. *Every discrete random source (X_t) with finite alphabet Σ and finite evolution dimension $d = e\dim(X_t)$ is asymptotically mean stationary.*

Proof. We sketch the argument whose details may be worked out in the standard way. The elements of the evolution space $\mathcal{E}$ can be identified with finite, signed measures on the σ-algebra generated by the cylinder sets relative to Σ^*. Since $\mathcal{E}$ is a finite-dimensional vector space, all norms on $\mathcal{E}$ are equivalent. In our discussion of pointwise convergence, we could therefore have chosen $||.||$ to be the norm of total variation, which implies (uniform) convergence on all elements of the underlying σ-algebra and thus the AMS property (see also Jacka and Roberts [11]).

$\diamond$

This assumption of finite evolution dimension is essential for Theorem 3 as not every discrete source with finite alphabet is (pointwise) asymptotically mean stationary. Consider, for example, the (deterministic) source (X_t) with binary alphabet $\Sigma = \{a, b\}$ that emits symbols in the following way. The source starts with $X_1 = a$. Then the source sends the other symbol b until the proportion of b's exceeds the threshold $2/3$. Then a sequence of a's follows until the proportion of a's exceeds $2/3$ *etc.*:

$$X = a\ bb\ aaa\ bbbbbb\ aaaaaaaaaaaa\ bb\ldots$$

The sampling frequency of any symbol $x \in \Sigma$ will fluctuate between $1/3$ and $2/3$ and never stabilize. Theorem 3 admits a weak converse, however.

Proposition 1. *Let (X_t) be a discrete AMS random source with stationary distribution $\overline{\mathbf{g}}$ and evolution space $\mathcal{E}$. Then*

$$\overline{\mathbf{g}} \in \mathcal{E} \quad \Longrightarrow \quad e\dim(X_t) < \infty.$$

Proof. Assume that the set $\mathcal{G} = \{\mathbf{g}^0, \mathbf{g}^1, \ldots\}$ of ground states of (X_t) is linearly independent (and hence a basis of $\mathcal{E}$) and suppose scalars α_i exist such that $\overline{\mathbf{g}} = \sum_{i=0}^{k} \alpha_i \mathbf{g}^i$. Then the application of the evolution operator yields

$$\overline{\mathbf{g}} = \psi(\overline{\mathbf{g}}) = \sum_{i=0}^{k} \alpha_i \psi(\mathbf{g}^i) = \sum_{i=0}^{k} \alpha_i \mathbf{g}^{i+1}\ ,$$

which contradicts the uniqueness of the representation of $\overline{\mathbf{g}} \in \mathcal{E}$ with respect to the basis $\mathcal{G}$. $\diamond$

Also a strong converse of Corollary 3 does not hold. To see this, consider the binary random source (X_t) with alphabet $\Sigma = \{a, b\}$ and an *independent* probability distribution in the following sense:

$$P\{X_1 = v_1, ..., X_t = v_t\} = P\{X_1 = v_1\} \cdot P\{X_2 = v_2\} \cdot \ldots \cdot P\{X_t = v_t\}.$$

Choose some $0 < p < 1$ and let

$$P\{X_t = a\} = p^t \quad \text{and} \quad P\{X_t = b\} = 1 - p^t.$$

The resulting process can then be shown to be AMS with infinite evolution dimension. (We refer to [15]) for the technical details.)

4 Markov Sources

We now describe a class of discrete random sources with finite evolution dimension that can be thought of as hidden Markov chains with possibly "negative transition probabilities".

Let $\Omega = \{1, \ldots, n\}$ be a finite set, which we imagine as the set of *hidden states* of some system $\mathcal{S}$. At each given time period $t \in \mathbb{N}$, $\mathcal{S}$ is in a definite (hidden) state $j \in \Omega$, which we observe in the next time period through the value $X_{t+1} = X(j)$ of the given *measuring function* $X : \Omega \to \Sigma$. Then $\mathcal{S}$ enters another state $i \in \Omega$ and emits the symbol $X_{t+2} = X(i)$ *etc.*

Setting $X_0 = \square$, we assume that the process (X_t) is a *linear* random process in the following sense: Once the string $w = w_1 \ldots w_t$ has been observed at time $t + 1$, our knowledge about the system is encoded in a vector $\pi_w \in \mathbb{R}^n$ such that

$$p(w) = \Pr\{X_1 = w_1, \ldots, X_t = w_t\} = \sum_{j=1}^{n} \pi_w(j) \ .$$

Moreover, with every $a \in \Sigma$, we assume a linear operator $S^a : \mathbb{R}^n \to \mathbb{R}^n$ to exist such that

$$\pi_{wa} = S^a \pi_w \quad \text{for all } w \in \Sigma^*.$$

Denoting the initial situation by $\pi^0 = \pi_\square$ and setting $M := \sum_{a\in\Sigma} S^a$, we have a situation as in Section 2.3 with the multilinear expansion

$$\pi^t := M^t \pi^0 = \big(\sum_{a\in\Sigma} S^a\big)^t \pi^0 = \sum_{w\in\Sigma^t} S^w \pi^0 = \sum_{w\in\Sigma^t} \pi_w = \sum_{w\in\Sigma^t} p(w)\pi'_w \ ,$$

where we write $\pi_w = p(w)\pi'_w$ to stress the formal analogy with Section 2.3 and call π'_w the *state vector* relative to w. Thus, if $p(w) \neq 0$, we find

$$S^a \pi'_w = \frac{p(wa)}{p(w)} \pi'_{wa} = p(a|w)\pi'_{wa}$$

and see that M always generates the expected next state vector

$$M\pi'_w = \sum_{a\in\Sigma} p(a|w)\pi'_{wa} \ .$$

The $(n \times n)$-matrix $M = [m_{ij}]$ can be interpreted as a generalized transition matrix for Ω with the column sum property

$$\sum_{i=1}^{n} m_{ij} = 1 \quad (j = 1, \ldots, n).$$

However, some of the "probabilities" m_{ij} might be negative.

REMARK. In the case of a hidden Markov chain (X_t) (see, *e.g.*, Elliot *et al.*[4]) we have fixed transition probabilities m_{ij} on the set Ω of hidden states, where

$$m_{ij} \geq 0 \quad \text{with} \quad \sum_{i=1}^{n} m_{ij} = 1 \quad (j = 1, \ldots, n),$$

and we start with a non-negative $\pi^0 \geq \mathbf{0}$. S^a is the matrix with row i equal to row i of the Markov transition matrix $\mu = [m_{ij}]$ if $X(i) = a$ and zero otherwise. Hence we arrive at the non-negative transition matrix

$$M = \sum_{a \in \Sigma} S^a = \mu \, .$$

Theorem 4. *A linear random source (X_t) with n hidden states has finite evolution dimension $e\dim(X_t) \leq n$ and hence is asymptotically mean stationary.*

Proof. Let $\mathcal{P} = [p(v|w)]$ be the prediction matrix and $\Pi' = [\pi'_w]$ the state vector matrix of (X_t). With the notation $\mathbf{1}^T = [1, 1, \ldots]$, we have for any $v \in \Sigma^*$,

$$p(v|w) = p(v|w)\mathbf{1}^T \pi'_{wv} = \mathbf{1}^T S^v \pi'_w \quad \text{for all } w \in \Sigma^*,$$

i.e., the v-row of $\mathcal{P}$ is a linear combination of the rows of Π'. So the linear dimension of the row space of $\mathcal{P}$ is bounded by the rank rk $\Pi' \leq n$. Consequently, also the linear dimension of the column space $\mathcal{V}$ is bounded by n and we have

$$e\dim(X_t) \leq \dim \mathcal{V} \leq n \, .$$

$\diamond$

REMARK. Our model of linear random sources with a finite number of hidden states is equivalent to the model of *finitary linearly dependent processes* (LDPs) (see Gilbert [6], Heller [8],[9] or Ito *et al.* [10]) or the model of *observable operator models* (OOMs) (see Jaeger [12]). It is not difficult to see that these are equivalent to the model of discrete random sources with finite-dimensional prediction space.

Jaeger [12] gives the example of a (finite-dimensional) OOM that cannot be presented as a classical finite-state hidden Markov chain, which exhibits the HMMs to form a proper subclass of the class of Markov sources described above. These in turn form a proper subclass of the processes with finite evolution dimension, as follows from our next result (and the example given in its proof).

Lemma 4. *There exists a stationary binary random source with infinite dimensional prediction space.*

Proof. Assume $\Sigma = \{a, b\}$ and let $\|v\|$ denote the Hamming norm on $\{a, b\}^*$, *i.e.*, the numbers of letters a occurring in the word $v \in \{a, b\}^*$. We now define a probability function $p : \Sigma^* \to \mathbb{R}$ by letting

$$p(v) = (t+1)^{-1} \binom{t}{\|v\|}^{-1} \quad \text{whenever } v \in \{a, b\}^t.$$

So the probability $p(v)$ of a word $v \in \{a,b\}^*$ depends only on its length $t = |v|$ and the number $\|v\|$ of a's occurring in it (but not on the particular order of their occurrence). In particular, the probability for a random word v of length $|v| = t$ to contain exactly m letters a is

$$P\{\|v\| = m\} = 1/(t+1) \quad \text{for } m = 0, 1, \ldots, t.$$

A straightforward computation shows

$$p(va) + p(vb) = p(av) + p(vb) = p(v) \quad \text{for all } v \in \{a,b\}^*.$$

So p yields a stationary binary random source (X_t). To see that the associated prediction space $\mathcal{V}$ is infinite-dimensional, consider the following $(m \times m)$-matrices A_m, where

$$A_m = \begin{bmatrix} p(\square) & p(a) & \cdots & p(a^{m-1}) \\ p(a) & p(aa) & \cdots & p(a^m) \\ & & \ddots & \\ \cdots & \cdots & & \cdots \\ p(a^{m-1}) & p(a^m) & \cdots & p(a^{2m-2)}) \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+1} \\ & & \ddots & \\ \cdots & \cdots & & \cdots \\ \frac{1}{m} & \frac{1}{m+1} & \cdots & \frac{1}{2m-1} \end{bmatrix}.$$

Induction on m easily exhibits these matrices to be regular and hence of full rank $\mathrm{rk}\, A_m = m$. It follows that the rank of the infinite subset $\{\mathbf{g}_\square, \mathbf{g}_a, \mathbf{g}_{aa}, \ldots\} \subseteq \mathcal{V}$ is not bounded, *i.e.*, $\mathcal{V}$ is an infinite-dimensional vector space.

$\diamond$

ACKNOWLEDGEMENT. The authors want to thank Herbert Jäger in particular and also the unknown reviewers for many helpful discussions and comments.

References

1. D. Aharonov, A. Ambainis, J. Kempe, U. Vazirani, "Quantum walks on graphs", in *Proc. of 33rd ACM STOC, New York*, 2001, pp. 50-59.
2. R.K. Brayton, C.H. Tong, "Stability of dynamical systems: A constructive approach", *IEEE Transactions on Circuits and Systems*, vol. 26, pp. 224-234, 1979.
3. S.P.M Choi, D.-Y. Yeung, N.L. Zhang, "Hidden-Markov decision processes for nonstationary sequential decision making", in: *Sequence Learning*, Lecture Notes in Artificial Intelligence 1828, R. Sun and C.L. Giles, Eds. Berlin: Springer-Verlag, 2000, pp. 264-287.
4. R.J. Elliot, L. Aggoun, J.B. Moore, *Hidden Markov Models*. Heidelberg: Springer-Verlag, 1995.
5. U. Faigle, A. Schönhuth, "Quantum predictor models", *Electonic Notes in Discrete Mathematics*, vol. 25, pp. 149-155, 2006.
6. E.J. Gilbert, "On the identifiability problem for functions of finite Markov chains", *Ann. Math. Stat.*, vol. 30, pp. 688-697, 1959.
7. T.S. Han and K. Kobayashi, *Mathematics of Information and Coding*. Providence, R.I.: Amer. Math. Soc. Mathematical Monographs 203, 2002.
8. A. Heller, "On stochastic processes derived from Markov chains", *Ann. Math. Statist.*, vol. 36, pp. 1286-1291, 1965.
9. A. Heller, "Probabilistic Automata and Stochastic Transformations", *Mathematical Systems Theory*, vol. 1(3), pp. 197-208, 1967.
10. H. Ito, S.-I. Amari, K. Kobayashi, "Identifiability of hidden Markov information sources and their minimum degrees of freedom", *IEEE Transactions on Information Theory*, vol. 38, pp. 324-333, 1992.

11. S.D. Jacka, G.O. Roberts, ”On strong forms of weak convergence”, *Stochastic Processes and Applications*, vol. 67, pp. 41-53, 1997.
12. H. Jaeger, ”Observable operator models for discrete stochastic time series”, *Neural Computing*, vol. 12, pp. 1371-1398, 2000.
13. R.M. Gray and J.C. Kieffer, “Asymptotically mean stationary measures”, *Annals of Probability*, vol. 8, pp. 962-973, 1980.
14. J.C. Kieffer, M. Rahe, ”Markov channels are asymptotically mean stationary”, *SIAM J. Math. Anal.*, vol. 12(3), pp. 293-305, 1981.
15. A. Schönhuth, ”Diskrete stochastische Vektorräume”, PhD thesis, University Cologne, Germany, 2006.